256/06
03.06.2022

Федеральное государственное
бюджетное учреждение науки
Институт проблем
передачи информации
им. А.А. Харкевича
Российской академии наук

ИППИ РАН

Большой Каретный пер., д. 19, стр. 1, Москва, 127051
ОКПО: 02699464 ОГРН: 1037700064940
ИНН/КПП: 7707020131/770701001
тел.: (495) 650-42-25 | факс: (495) 650-05-79 | director@iitp.ru

_____ 20____ г. №11615-_____ / _____

На № _____ от _____

УТВЕРЖДАЮ
Директор ИППИ РАН им. А. А. Харкевича
д.ф.-м.н., профессор РАН
_____ А. Н. Соболевский
«20» _____ 2022 г.

**ОТЗЫВ**
**ведущей организации на диссертационную работу**
***Зиновьевой Анастасии Юрьевны***
**«Модель многоязычного интеллектуального контент-анализа (на материале франко-, англо- и русскоязычных новостных сообщений о террористической деятельности)», представленную на соискание ученой степени кандидата филологических наук по специальности 10.02.21 – прикладная и математическая лингвистика**

Диссертационная работа Анастасии Юрьевны Зиновьевой посвящена разработке методов интеллектуального многоязычного контент-анализа и применению разработанных методов к предметной области «Терроризм». Многоязычный интеллектуальный контент-анализ определяется как извлечение из неструктурированных разноязычных корпусов текстов ограниченной предметной области контента, релевантного информационному запросу пользователя, его числовую обработку, интерпретацию и представление в удобной для пользователя форме.

**Актуальность проведенного исследования** несомненна. Она вытекает хотя бы из того, что проблематика излечения смысла из потока текстов и его интеллектуальный анализ, хотя и привлекла большое внимание исследовательского сообщества в последнее

десятилетие, но все же остается полной лакун. До сих пор отсутствует полная и научно-обоснованная методология построения систем контент-анализа, способная удовлетворить потребности общества в автоматизированной интерпретации содержания текстов. В построение такой методологии и вносит заметный вклад диссертация А. Ю. Зиновьевой. Особо следует отметить выбор предметной области для апробации разработанного метода. Трудно подобрать более актуальную тематику для анализа новостного потока, чем терроризм.

Разработанная автором методология описана во всех деталях и **готова к использованию на практике**. В первую очередь это касается базы знаний предметной области «Терроризм», которая в значительной степени доведена до уровня практической применимости и использование которой позволит частично автоматизировать и повысить оперативность аналитической деятельности в сфере политического прогнозирования и контртерроризма. Однако этим значимость полученных результатов не исчерпывается. Диссертация продвигает вперед **фундаментальную задачу** семантического анализа неструктурированных текстов, включая выявление имплицитных компонентов значения, которые продолжают оставаться на периферии внимания лингвистов. Для охвата этого слоя содержания текста автору приходится значительно расширить традиционный инструментарий лингвистического анализа, включив в него онтологическое моделирование и аксиомы логического вывода. Следует рекомендовать включить соответствующие разделы в программу подготовки как теоретических, так и компьютерных лингвистов.

**Тема и содержание** представленной на рецензию диссертации А.Ю. Зиновьевой полностью **соответствуют паспорту** научной специальности 10.02.21 – Прикладная и математическая лингвистика – в части разработки языковедческой теории на основе изучения специфических современных практических задач как собственно лингвистики, так и отдельных прикладных областей, таких как информационный поиск, компьютерная лексикология, лексикография и др. Соответствие содержания диссертационного исследования специальности подтверждается также апробацией работы на 7 международных, всероссийских и вузовских научных конференциях.

Основные научные результаты, выводы и рекомендации диссертационного исследования **отражены** в 10 работах, 4 из которых опубликованы в рецензируемых

изданиях, включенных в Перечень ВАК, 3 – в БД WoS, что **соответствует пп.11, 12 и 13 Положения о присуждении ученых степеней.** Опубликованные работы и автореферат отражают содержание работы с достаточной полнотой. Представленная диссертация оформлена в соответствии с требованиями, предъявляемыми к работам данного жанра.

Переходя непосредственно к результатам, полученным в диссертации, перечислим несколько особенно значимых моментов.

1. Ориентация модели на ограниченную предметную область и ограниченный подъязык позволяет заметно повысить качество семантического анализа текста. Она облегчает задачу семантической дизамбигуации лексики и грамматических конструкций и обеспечивает учет фоновых знаний пользователя, которые являются мощным и еще недостаточно оцененным источником для извлечения имплицитных компонентов значения.
2. Анализ новостных сообщений предметной области проведен на материале соизмеримых корпусов текстов на трех языках – русском, английском и французском. Многоязычность методики не только значительно расширяет непосредственную эффективность контент-анализа, но и позволяет достичь языковой независимости базы знаний, являющейся ключевым компонентом принятого автором подхода к моделированию многоязычного интеллектуального контент-анализа.
3. В качестве основного метода интеллектуального контент-анализа выбрано представление концептуальной структуры текстов. Определение концептуальной структуры основано на онтологическом анализе с использованием специально построенной многоязычной предметно-ориентированной лингвистической онтологии.
4. Архитектура модели контент-анализа состоит из двух частей: лексико-онтологической базы знаний и алгоритмических процедур интеллектуального контент-анализа. Из всех компонентов модели база знаний обладает наиболее разветвленной структурой и наиболее сложна для построения. В нее входят следующие элементы:

   – многоязычная онтология предметной области «Терроризм»;

   – база экземпляров онтологии;

– онтолексиконы и ономастиконы, содержащие релевантные для предметной области лексемы;

– правила онтологического анализа и логического вывода;

– правила формирования концептуально-лексических фреймов и оформления результатов работы модели.

5. Разработана единая трехэтапная методика анализа текстов, применявшаяся ко всем трем рабочим языкам. Она позволила выделить специфические характеристики подъязыка предметной области, которые могут быть использованы при моделировании многоязычного интеллектуального контент-анализа. Следует отметить разработанную автором прескриптивно-дескриптивную методику отбора единиц, отражающих содержание предметной области.
6. В результате компонентного и контекстного анализа лексических единиц было образовано 25 концептуальных классов.
7. Были обнаружены и описаны явления концептуальной однозначности, концептуальной неоднозначности и концептуального синкретизма, представляющие интерес с точки зрения теоретической семантики.
8. Проведен анализ подъязыка новостных сообщений предметной области. Обнаружено, что на уровне структуры текста проявляются ограничения подъязыка, обусловленные жанром новостных сообщений. Новостные сообщения предметной области «Терроризм» в подавляющем большинстве случаев во всех рассмотренных языках имеют структуру релевантности типа «перевернутая пирамида». В такой структуре наиболее релевантная информация приводится в заголовке, и релевантность дальнейшей информации последовательно уменьшается.
9. Ограничения и закономерности, заданные предметной областью, проявляются и на лексико-семантическом уровне. Бóльшая часть контента в рассмотренных языках отражена в существительных, причем конкретные пропорции частотности уникальных лексем почти совпадают, в то время как пропорции частотности словоупотреблений различаются.

Полученные в диссертации результаты представлены последовательно и изложены хорошим языком.

Вместе с тем хотелось бы высказать некоторые мелкие замечания, касающиеся аксиом онтологии предметной области.

1. На стр. 114 приводится Аксиома 2

   «$\forall\ T\ instrument(T, Cc) \rightarrow type(T, Taa)$, где T — TERROR-ATTACK, Cc — GUN, Taa — GUN-ATTACK; т. е. для всех терактов верно, что если теракт был совершен с использованием огнестрельного оружия, то типом теракта является стрельба». Далее автор делает замечание: «В этой аксиоме концепт GUN может быть заменен на любой дочерний концепт WEAPON, а концепт GUN-ATTACK — на любой дочерний концепт TERROR-ATTACK соответственно». Это замечание неточно. Не любой дочерний концепт Weapon совместим с произвольным дочерним элементом Terror-attack. Например, Knife не может быть инструментом теракта типа CyberAttack. Впрочем, в списке аксиом, приводимом в Приложении E, корреляция между инструментами и типами терактов сформулирована корректно.
2. Возможно, стоило бы сформулировать и обратные аксиомы типа: «если типом теракта является стрельба, то в качестве инструмента было использовано огнестрельное оружие».
3. Некоторые аксиомы из приложения E являются следствиями из других аксиом и могут быть устранены из списка. Например,

Аксиома 1. $\forall\ T\ located(T, La) \cap part\ of(La, Lc) \rightarrow located(T, Lc)$,
где T — TERROR-ATTACK, La — CITY, Lc — COUNTRY; т. е. для всех терактов верно, что, если теракт произошел в некоем городе и известно, что этот город находится в некой стране, то, следовательно, теракт произошел в этой стране.

Аксиома 2. $\forall\ T\ located(T, Laa) \cap part\ of(Laa, Lc) \rightarrow located(T, Lc)$,
где T — TERROR-ATTACK, Laa — CAPITAL-CITY, Lc — COUNTRY; т. е. для всех терактов верно, что, если теракт произошел в некой столице и известно, что эта столица находится в некой стране, то, следовательно, теракт произошел в этой стране.

Поскольку известно, что CAPITAL-CITY является подклассом CITY, то аксиома 2 вытекает из аксиомы 1.

Высказанные замечания ни в коей мере не снижают теоретической и практической значимости проведенного исследования. В целом, диссертационное исследование выполнено на очень высоком уровне. Работа Анастасии Юрьевны Зиновьевой «Модель многоязычного интеллектуального контент-анализа (на материале франко-, англо- и

русскоязычных новостных сообщений о террористической деятельности)» является оригинальным, самостоятельным и завершенным исследованием, отвечает требованиям пп. 9-14 Положения о присуждении ученых степеней, предъявляемым к кандидатским диссертациям (утвержденным Постановлением Правительства РФ от 24.09.2013 №842), а ее автор Зиновьева Анастасия Юрьевна заслуживает присуждения ученой степени кандидата филологических наук по специальности 10.02.21 – прикладная и математическая лингвистика.

Диссертация и автореферат заслушаны и обсуждены на заседании научного семинара Лаборатории компьютерной лингвистики Института проблем передачи информации РАН 20 мая 2022 г., протокол № 3.

Отзыв составил:

главный научный сотрудник Лаборатории компьютерной лингвистики,

доктор филологических наук, профессор

Богуславский Игорь Михайлович

Подпись Богуславского И.М.
УДОСТОВЕРЯЮ
Зав. канцелярией